Hierarchical Segmentation Using Tree-Based Shape Spaces

Yongchao Xu, Edwin Carlinet, Thierry Géraud, Laurent Najman

Hierarchical Segmentation Using Tree-Based Shape Spaces

Yongchao Xu, Edwin Carlinet, Thierry Géraud, Laurent Najman

Abstract—Current trends in image segmentation are to compute a hierarchy of image segmentations from fine to coarse. A classical approach to obtain a single meaningful image partition from a given hierarchy is to cut it in an optimal way, following the seminal approach of the scale-set theory. While interesting in many cases, the resulting segmentation, being a non-horizontal cut, is limited by the structure of the hierarchy. In this paper, we propose a novel approach that acts by transforming an input hierarchy into a new saliency map. It relies on the notion of shape space: a graph representation of a set of regions extracted from the image. Each region is characterized with an attribute describing it. We weigh the boundaries of a subset of meaningful regions (local minima) in the shape space by extinction values based on the attribute. This extinction-based saliency map represents a new hierarchy of segmentations highlighting regions having some specific characteristics. Each threshold of this map represents a segmentation which is generally different from any cut of the original hierarchy. This new approach thus enlarges the set of possible partition results that can be extracted from a given hierarchy. Qualitative and quantitative illustrations demonstrate the usefulness of the proposed method.

Index Terms—Graph, shape space, tree of shapes, minimum spanning tree, α-tree, binary partition tree, object spotting, image segmentation, hierarchy, hierarchical segmentation, saliency map.

1 Introduction

Image segmentation is one of the oldest and most challenging problems in image processing. As shown in [1], even for a human observer, it is hard to determine a unique meaningful segmentation of a given image f. As promoted by Guigues *et al.* in [2], a low level segmentation tool should remain scale uncommitted, because the structures which can be useful to high level task can have arbitrary size. In other words, a segmentation should output a multi-scale description of the image f. An usual approach to overcome the difficulty of finding a unique meaningful partition, and to satisfy the multi-scale property, is to compute a hierarchy of segmentations, which encodes a set of segmentations from fine to coarse. Such a hierarchy is usually represented by a dendrogram as shown in Fig. 2 and Fig. 3. Another representation is the saliency map, originally introduced in [3], and independently rediscovered by Guigues *et al.* under the name of contour disappearance map in [2]. Recently, associated to a specific learning-based algorithm, this representation has been popularized under the name of ultrametric contour map [1]. Each threshold of the saliency map gives a segmentation result, and conversely, by stacking a set of segmentations satisfying a hierarchical property, one obtains a saliency map. A theoretical study and several characterizations of saliency maps can be found in [4]; efficient algorithms to compute saliency maps are given in [5].

In this paper, we present a novel general framework for obtaining meaningful hierarchical partitions from any hierarchical representation of an image. The proposed framework is the counterpart of the shape-space filtering framework introduced in [6]. Recall that the core idea of this later framework, as shown in Fig. 1, is to apply some morphological operators to the shape graph-space of connected components extracted from the image: the vertices of the shape graph-space are the connected components, while the edges are provided thanks to the parenthood relationship between the connected components (*i.e.*, the neighbors of a vertex are its children and its parent). The shape-space filters proposed in [6] are connected operators [7, 8], filtering out from the shape space (and hence from the image) the unwanted connected components. In this paper, we weigh the vertices of the shape space with an attribute describing the characteristics of the regions, and we consider all the local minima of the weighted shape graph-space as candidate regions of the final meaningful partition. We then obtain a saliency map $\mathcal{M}_\mathcal{E}$ by weighing the boundaries of the candidate regions by their importance, measured by the extinction values [9] of the corresponding local minimum. Meaningful partitions can be obtained by thresholding this new saliency map $\mathcal{M}_\mathcal{E}$. In particular, if the attribute characterizes the shape (*e.g.*, circularity, upper triangularity, etc.) of each region in the shape space, the obtained saliency map $\mathcal{M}_\mathcal{E}$ represents a shape-oriented hierarchical image segmentation: in such a representation, the regions with the desired shapes

- *Yongchao Xu, Edwin Carlinet, and Thierry Géraud are with EPITA Research and Development Laboratory (LRDE), 14-16 rue Voltaire, FR-94270 Le Kremlin-Bicêtre, France.*
- *Yongchao Xu, Edwin Carlinet, Thierry Géraud, and Laurent Najman are with the Laboratoire d'Informatique Gaspard-Monge, Université Paris-Est, Équipe A3SI, ESIEE Paris, 93160 Noisy-le-Grand, France.*
- *Yongchao Xu is with LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France*
  E-mails: {yongchao.xu, thierry.geraud}@lrde.epita.fr, l.najman@esiee.fr

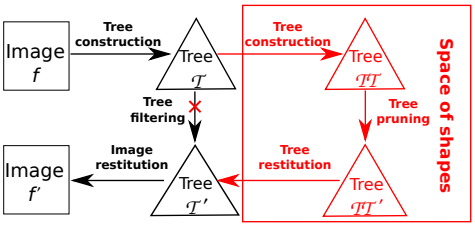
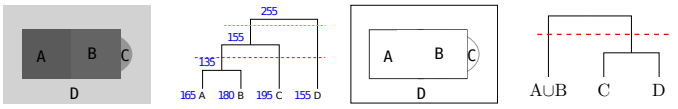


Fig. 1: Shape-space filtering framework: we replace the black path with the red path, thus extending the classical connected operator framework (black path).



(a) Image f (b) Input H (c) $\mathcal{M}_\mathcal{E}$ (d) Output H'

Fig. 2: A synthetic example of a hierarchy transformation through $\mathcal{M}_\mathcal{E}$. Blue numbers are the attribute values (*e.g.*, inverse of average of gradient's magnitude along the contour). For visualization purpose, the inverse of $\mathcal{M}_\mathcal{E}$ is shown. A cut (the red dashed line) of this new hierarchy H' is different from any cut of H (*e.g.*, the red cut or the green cut in (b)).

are brought to the fore. This scheme can be extended to highlight the objects with any specific characteristics described by the attribute.

The proposed framework to obtain extinction-based saliency maps $\mathcal{M}_\mathcal{E}$ has several interesting consequences. First of all, it allows us to transform any hierarchical representation of the image into a hierarchy of image segmentations. Secondly, if the input hierarchical representation is a hierarchy of image segmentations H, the partitions obtained by thresholding the proposed extinction-based saliency maps $\mathcal{M}_\mathcal{E}$ can be different from any cut of the initial hierarchy H. This last point was first highlighted in a preliminary version of this study [10]. Consequently, as illustrated in the synthetic example of Fig. 2, our method potentially modifies the structure of H. This might give interesting results in many cases: indeed, as the set of partitions that can be extracted from a shape space contains the set of non-horizontal cuts that can be extracted from the hierarchy, we have a broader choice when selecting a partition from the shape space. Furthermore, the obtained hierarchy of segmentations highlights the regions of specific characteristic represented by the attribute.

The rest of this paper is organized as follows. In Section 2, we review a number of background materials related to our proposal. Section 3 compares the proposed framework *w.r.t.* some closely related methods. The methodology of the proposed framework for hierarchical segmentations is detailed in Section 4. In Section 5, we depict three experimental results using the proposed framework with different input hierarchies and attributes characterizing the regions: 1) the α-tree [11] and an attribute inspired from the work in [12] to extend the α-tree for generic image segmentation; 2) the tree of shapes and shape attributes for traffic sign detection; 3) the tree of shapes and a specifically designed attribute for document extraction in videos captured by smartphones. Finally we conclude in Section 6.

2 Background

The hierarchical representations of the image can be classified into two families. The first one is the family of hierarchy of segmentations reviewed in Section 2.1. Partitions are usually obtained via hierarchical cuts presented in Section 2.2. The second family is the set of threshold decomposition-based trees reviewed in Section 2.3. The common features of these hierarchies lead to the notion of shape space that we have introduced in [13, 6], and which is shortly presented in Section 2.4.

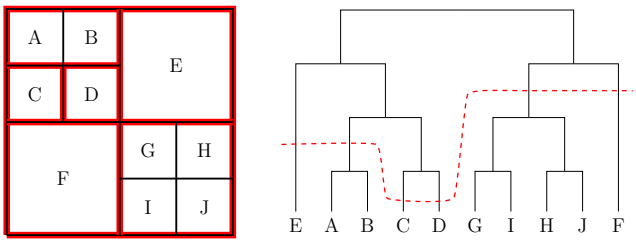
In this paper, an image is modeled as a graph $G = (V, E)$ weighted by a function f, where V is the image domain (*i.e.*, the set of pixels), $E \subset V \times V$ is the set of edges, and $f : V \to \mathbb{R}$ represents the image intensity; in other words, $f(v)$ is the grey-level of the pixel v.

2.1 Hierarchy of image segmentations

A *partition* or a *segmentation* of a set V is a collection of subsets R_i (*i.e.* region) of V such that V is the disjoint union of the subsets. Very often, each one of the regions of a segmentation is connected for the underlying graph $G = (V, E)$. A segmentation P_i is *finer* than the partition P_k if any region of P_i is a subset of a region of P_k. A hierarchy of segmentations H is a chain of nested partitions P_i: $H = \{P_i \,|\, 0 \leq i \leq n, \forall j, k,\ 0 \leq j \leq k \leq n \Rightarrow P_j \sqsubseteq P_k\}$, where $P_j \sqsubseteq P_k$ denotes that the partition P_j is finer than the partition P_k. We denote by P_n the partition $\{V\}$ which segments the entire image as a single region, and by P_0 the finest partition of the graph (V, E). In other words, a hierarchy of segmentations H is a set of regions, such that: 1) $\{V\} \in H$; (2): for each region $R \in P_0$, $R \in H$; (3): for each pair of distinct regions (R, R'), where $R \in H, R' \in H$, $R \cap R' \neq \emptyset \Rightarrow R \subset R'$ or $R' \subset R$. Relation (3) formalizes that two distinct regions in a hierarchy of segmentations are either disjoint or nested.

A hierarchy of image segmentations can be represented via a special type of tree called the *dendrogram*. The root of the tree represents the entire image, and the leaves are the regions of the finest partition P_0, while an intermediary node R represents the merging of regions represented by the nodes just below, known as the children of the parent R. Examples of a hierarchy of segmentations represented by a dendrogram are shown in Fig. 2 and Fig. 3.

An example of hierarchy of segmentations is the α-tree [11], also known as the hierarchy of constrained connectivity [14]. It relies on the notion of α-connectivity [14]. For a pair of neighboring pixels $p \in V$ and $q \in V$, let $d(p, q)$ be the dissimilarity measure between p and q. Then two pixels p and q are said to be α-connected if there is a path between p and q



(a) Original image (b) A cut of H

Fig. 3: A synthetic image (a), and its associated dendrogram (b) representing a hierarchy of image segmentations H. The red dashed curve in (b) is a cut of the dendrogram (*i.e.*, hierarchy), the partition given by this cut is illustrated with red boundaries in (a).

such that for any pair of adjacent pixels (p_i, p_{i+1}) in the path, $d(p_i, p_{i+1}) \leq \alpha$ always holds. Based on this notion of α-connectivity, an α-connected component [14] containing a pixel p is defined as α-CC$(p) = \{p\} \cup \{q \mid p$ and q are α-connected$\}$. There exists an inclusion relationship between the α-connected components: $\forall\, x \in V$, $\alpha_1 \leq \alpha_2 \Rightarrow \alpha_1$-CC$(x) \subseteq \alpha_2$-CC$(x)$, which yields the α-tree [11]. It has been shown in [15] that an α-tree is equivalent to the Min-tree [7] of a minimum spanning tree (MST) [16], which provides an efficient algorithm to compute the α-tree. The interested reader can refer to [14, 11] for more details about the α-tree.

2.2 Cuts in a hierarchy

A hierarchy of segmentations H generates a subset of all possible partitions of image. Selecting a "best" or optimal one is usually achieved by the notion of cuts [2]. A cut of a hierarchy H is a subset of H which intersects any path from the base to the top of H exactly once. Equivalently, a cut is a partition P whose regions are taken from the regions represented by nodes in H. An example is shown in Fig. 3.

For an indexed hierarchy which is a pair (H, λ), where λ is a positive function (*e.g.*, the scale), defined on H such that for two nested regions $R, R' \in H, R \subset R'$, we have $\lambda(R) < \lambda(R')$, the simplest cut is an horizontal cut, *i.e.*, threshold the hierarchy H based on its associated index λ. An example is the work in [14], where the author proposes to threshold the local range α or a global range ω of the α-tree [11]. This corresponds to cut horizontally the α-tree by setting the index to respectively α or ω.

A more evolved hierarchy cut consists in minimizing an energy functional subordinated to the hierarchy H. A popular example is the work in [2], where the authors propose the scale-set theory by considering a rather general formulation of segmentation problem. It involves minimizing a two-term-based energy functional that can be written as $E_{\lambda_s} = \sum_{R \in P} \lambda_s C(R) + D(R)$ for a partition P, where C is a decreasing regularization term, D is an increasing goodness-of-fit term, and λ_s is a parameter. This energy E_{λ_s} is called multi-scale affine separable energy. In [2], the authors use dynamic programming to efficiently compute two scale parameters λ_s^+ and λ_s^- for each region $R \in H$, where λ_s^+ (*resp.* λ_s^-) corresponds to the smallest parameter λ_s such that the region $R \in H$ belongs (*resp.* no longer belongs) to the optimal segmentation by minimizing E_{λ_s} subordinated to H. There may exist some regions R such that $\lambda_s^-(R) \leq \lambda_s^+(R)$, which implies that the region $R \in H$ does not belong to any optimal cut of H by minimizing the energy E_{λ_s}. One removes these regions from the hierarchy H and updates the parenthood relationship (for a removed region R, the parent of its children is set to the parent of the region R), which yields a hierarchy H'. Then for any given λ_s, the optimal λ_s-cut of the original hierarchy (H, λ) by minimizing E_{λ_s} is given by thresholding λ_s^+ of H'.

The optimal cuts of a hierarchy of segmentations by minimizing the energy E_{λ_s} in [2] has been extended by Serra and Kiran [17] via the notion of h-increasingness. They propose a new approach to find optimal cuts of H in one pass based on energy minimization. The interested reader can refer to [17, 18, 19] for more details.

Another interesting cut of a hierarchy of segmentations is proposed by Cardelino *et al.* [20]. The method is based on the use of the *a contrario* model [21]. They assign a meaningfulness reflected by the number of false alarms (NFA) to each possible partition spanned by that hierarchy. The optimal partition is simply given by the most meaningful one.

2.3 Threshold decomposition-based trees

Another type of hierarchical representations is based on threshold decomposition. The simplest one is the Max-tree or its dual tree called the Min-tree [7]. They are based on the inclusion relationship of connected components of respectively upper level sets and lower level sets. For any $\lambda \in \mathbb{R}$, the upper level sets $\mathcal{X}_\lambda$ and lower level sets $\mathcal{X}^\lambda$ of an image f are defined by $\mathcal{X}_\lambda(f) = \{v \in V \mid f(v) \geq \lambda\}$ and respectively $\mathcal{X}^\lambda(f) = \{v \in V \mid f(v) < \lambda\}$. Indeed, both upper and lower level sets have a natural inclusion structure: $\forall\, \lambda \leq \mu$, $\mathcal{X}_\lambda \supseteq \mathcal{X}_\mu$ and $\mathcal{X}^\lambda \subseteq \mathcal{X}^\mu$, which leads to two distinct and dual representations of an image, respectively the Max-tree and the Min-tree. The fusion of the Max-tree and Min-tree gives the tree of shapes [22], known also as the topographic map [23] through the notion of *shape*. A shape in the tree of shapes is defined as a connected component of an upper or a lower level set with its holes filled in.

A major difference between these threshold decomposition-based trees $\mathcal{T}_t$ and the hierarchies of segmentations $\mathcal{T}_h$ reviewed in Section 2.1 is that any cut of a type $\mathcal{T}_h$ gives a partition of the image domain, whereas any cut (except the root) of a type $\mathcal{T}_t$ yields a subset of the image domain.

2.4 Shape spaces

Both the hierarchies of segmentations reviewed in Section 2.1 and the threshold decomposition-based trees

reviewed in Section 2.3 have a tree structure. Each representation is composed of a set of connected components $\mathbb{C}$. Any two different elements $C_i \in \mathbb{C}, C_j \in \mathbb{C}$ are either disjoint or nested: $\forall C_i \in \mathbb{C},\ C_j \in \mathbb{C}, C_i \cap C_j \neq \emptyset \Rightarrow C_i \subseteq C_j$ or $C_j \subseteq C_i$. This property leads to the definition of *tree-based shape space* in [6]: a graph representation $G_{\mathbb{C}} = (\mathbb{C}, E_{\mathbb{C}})$, where each node of the graph represents a connected component in the tree, and the edges $E_{\mathbb{C}}$ are given by the inclusion relationship between connected components in $\mathbb{C}$.

It has been shown that the shape space has several interesting features. Firstly, It is an equivalent representation of an image, in the sense that the image can be reconstructed form the shape space. This representation inherently embeds a morphological scale space satisfying the *principle of causality* [24]. Furthermore, some of them are invariant to contrast changes and covariant to continuous (topological) transformations. Besides, contrary to scale-space, the contours of a given shape (connected component) correspond to actual contours in the image, without any blurring due to convolution with a kernel in the case of classical scale-space.

The shape space has been proved to be very useful in many applications. Examples are meaningful level lines selection [25], classification of images [26], texture indexing [27], scenery image analysis [28], image simplification and segmentation [7, 29, 30, 12, 31, 32, 14, 33], object detection [34, 35], and local feature detection [36, 13].

3 Comparison with related work

We focus on the comparison with two families of closely related work: shape-space filtering [6] in Section 3.1 and hierarchical cuts [2, 17, 18, 19, 20] in Section 3.2.

3.1 Comparison with shape-space filtering

The shape-space filters proposed in [6] are connected operators that filter out unwanted shapes from the shape space (hence from the image). The second tree representation $\mathcal{TT}$ (see Fig. 1) is constructed to perform the filtering. In this paper, we use the shape space and the second tree (see Fig. 4) in a different way. We consider all the local minima of the weighted shape space as candidate regions of the final partitions. We then rely on a second tree $\mathcal{TT}$ (being a Min-tree) to compute the extinction value [9] for each local minimum, which measures the importance of each underlying candidate region. The saliency map is obtained by weighing the extinction values to the boundaries of the corresponding candidate regions. This saliency map represents a hierarchy of image segmentations.

Another difference *w.r.t.* our previous work in [6] is on the construction of the input tree $\mathcal{T}$. In [6], the input tree is usually a threshold decomposition-based tree, where the input image is seen as a node-weighted graph. Yet, in this paper, the input tree can be any hierarchical representation of the image. For example, we use an α-tree in Section 5.1 where the input image is seen as an edge-weighted graph.

3.2 Comparison with hierarchical cuts

The cut-based methods [2, 17, 18, 19, 20] reviewed in Section 2.2 consist of cutting the input hierarchy of segmentations horizontally or non-horizontally to obtain a partition. They do not change the structure of the input hierarchy of segmentations H in the sense that for each region in H, either all its child regions or none of these two regions belong to the final partition. For example, in the hierarchy illustrated in Fig. 2 (b), either both $A \cup B$ and C (children of $A \cup B \cup C$) belong to the final partition (*e.g.*, the red cut), or none of these two regions belong to the final partition (*e.g.*, the green cut).

Our extinction-based saliency map $\mathcal{M}_{\mathcal{E}}$ is different from the cut-based approaches. Indeed, it does not follow the idea of cutting directly the hierarchy to obtain meaningful segmentations. As described in Section 4.3, we propose to consider a subset of regions in the hierarchy H as candidate regions of the meaningful partitions. Then we weigh their boundaries by their meaningfulness measured by the extinction values, which produces a saliency map $\mathcal{M}_{\mathcal{E}}$. Then segmentation results can be obtained by thresholding this saliency map. Note that a segmentation given by a cut of $\mathcal{M}_{\mathcal{E}}$ is usually different from any cut of the original shape space. For instance, in the original hierarchy H depicted in Fig. 2 (b), the region C merges with the region $A \cup B$, the region $C \cup D$ will never be a single region of a partition given by any cut of the hierarchy H. In contrast, the region $C \cup D$ is a possible single region of a partition given by thresholding extinction-based saliency map $\mathcal{M}_{\mathcal{E}}$ as depicted in Fig. 2 (c) and Fig. 2 (d).

Another difference is that our proposed framework also works for the threshold decomposition-based trees $\mathcal{T}_t$. Whereas, any cut (except cutting the root node) of $\mathcal{T}_t$ yields a subset of the image, not a partition of the image domain. The choice between a tree $\mathcal{T}_t$ or a hierarchy of segmentations $\mathcal{T}_h$ to construct the shape space depends on the application, the user has only to ensure that the regions of interest are present in the shape space.

Another advantage of our proposed framework is that it is not limited to the h-increasing attributes used in the state-of-the-art hierarchy transformation methods [3, 2, 14, 4, 17, 18, 19, 37]. We can use any attribute in our proposed framework. For example, if we have some prior knowledge (*e.g.*, the shape or some specific characteristics) about the regions of interest, we can integrate this information into the attribute. In consequence, the regions of interest are highlighted in the extinction-based saliency map.

4 Hierarchical segmentation based on shape spaces

In this section, we present our proposed framework of hierarchical segmentation using shape spaces. We start with a general overview of the proposed scheme in Section 4.1. Then we detail in Section 4.2 the basis of the framework: object spotting using the shape space

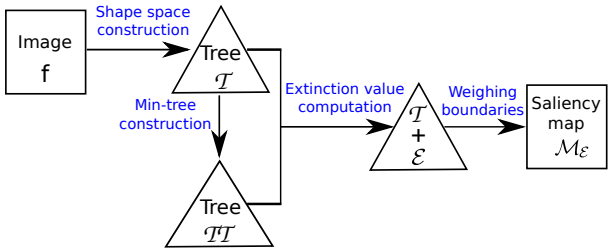


Fig. 4: General overview of the proposed scheme: shape space construction by a hierarchical representation of the image, followed by the computation of extinction values $\mathcal{E}$ via a Min-tree $\mathcal{TT}$ constructed on the shape space. Finally, the saliency map $\mathcal{M}_\mathcal{E}$ is obtained by weighing the boundaries of candidate regions by the extinction values of the corresponding local minima.

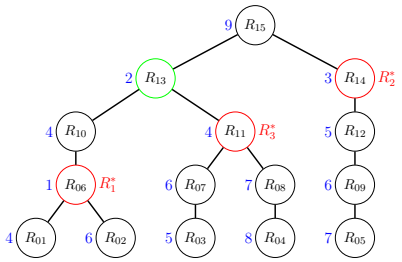


Fig. 5: Object spotting by repeating the process of selecting the "most likely" node R_i^* and discarding the ancestors and descendants of node R_i^*. Blue values are corresponding attribute values $\mathcal{A}$. The three nodes represented by red circles are the spotted objects, where R_1^* is spotted firstly, then R_2^* and R_3^*. Note that the green node R_{13} is more meaningful than R_{11}, but it is not spotted.

weighted by an attribute $\mathcal{A}$. The extension of this idea to produce an extinction-based saliency map $\mathcal{M}_\mathcal{E}$ is presented in Section 4.3 .

4.1 General overview of the proposed framework

The shape space that we have introduced in [6] (shortly reviewed in Section 2.4) is composed of a reasonable number of regions from fine to coarse, which provides a tremendously reduced search space for object spotting and segmentation. This motivates us to perform object spotting and segmentation tasks based on the shape space. The work of Vilaplana *et al.* [34] is such an instance, using the binary partition tree [29] to construct the shape space. Instead of being satisfied with a single spotting/segmentation result, we propose to extend the principle to hierarchical object segmentation. The general overview of the proposed scheme is depicted in Fig. 4.

The basic idea is to consider the local minima of the node-weighted shape space $(\mathbb{C}, E_\mathbb{C}, \mathcal{A})$ as candidate meaningful objects, and to extract only the significant local minima as the final spotted objects. We propose to use the extinction value [9] of each local minimum to characterize the significance of the underlying candidate object. The extinction values for all the local minima of the shape space $(\mathbb{C}, E_\mathbb{C}, \mathcal{A})$ can be efficiently computed via a Min-tree ($\mathcal{TT}$ in Fig. 4) constructed on $(\mathbb{C}, E_\mathbb{C}, \mathcal{A})$. Finally, we propose to weigh the boundaries of the candidate regions by their corresponding extinction values. This yields a new saliency map $\mathcal{M}_\mathcal{E}$ representing a hierarchy of image segmentations.

4.2 Object spotting using shape space

Suppose that the interesting objects that one would like to spot are present in the shape space constructed somehow, then the object segmentation problem is now reduced to how to retrieve them from the shape space. To achieve this, we assign to each region an attribute $\mathcal{A}$ capturing its characteristics. Examples of $\mathcal{A}$ are the compactness and the number of false alarms (NFA) measuring the meaningfulness of a region boundary proposed in [25]. If we have some prior information (*e.g.*, shape or color) about the interesting objects to be spotted, the attribute $\mathcal{A}$ can also be some specifically designed assessment measuring how much a region fits the prior knowledge.

Once the shape space is built and the attribute $\mathcal{A}$ is available, the object segmentation task is achieved by object spotting in the shape space (*i.e.*, search space). The most trivial spotting strategy is to choose the "most likely" one. It is useful if there is only one interesting object in the image. However, in most cases, the number of interesting objects is unknown, and is usually more than one. In this case, one possibility is to first of all spot the "most likely" region R_1^* among all the regions in the shape space, and discard all the ancestors and descendants of R_1^*. Then retrieve a second "most likely" region R_2^* among the remaining regions in the shape space, and discard again its descendants and ancestors. This selecting and discarding process is repeated until all the regions are either spotted or discarded. In consequence, a set of regions $\{R_i^*, | i = 1, \ldots, n\}$ will be spotted, where the number of spotted objects n is decided by the algorithm. Such an object spotting process is depicted in Fig. 5. This spotting strategy might give interesting results in some cases, but it ignores the fact that several interesting objects may be present in a same branch of $\mathcal{T}$, which means one may be included in another. For example the region R_{06} and R_{13} in Fig. 5.

The notion of "most likely" is usually modeled by the extremum of the attribute $\mathcal{A}$. Following this idea, we propose to spot the local minima of the attribute $\mathcal{A}$ as interesting objects. This strategy enables to spot different regions in the same branch. For instance, in Fig. 5, the three local minima of the attribute $\mathcal{A}$ in the shape space are respectively R_{06}, R_{13}, and R_{14}, where nodes R_{06} and R_{13} are in the same branch.

There are usually many local minima of a given attribute $\mathcal{A}$ on any hierarchical representation of the image. An example of an attribute $\mathcal{A}$ based on the meaningfulness of the region boundaries is depicted in

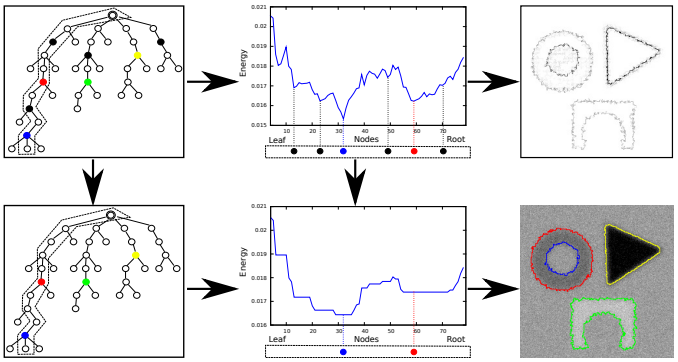


Fig. 6: An example of the object spotting scheme by selecting important local minima of an attribute $\mathcal{A}$ as meaningful objects. The shape space is built from the tree of shapes, and $\mathcal{A}$ is the context-based energy estimator [35]. Left: Tree weighted by $\mathcal{A}$ (top) and filtered $\mathcal{A}'$ (bottom); Filled circles: local minima; Colorized filled circles: resistant local minima after connected filtering in the shape space. Middle: Evolution of $\mathcal{A}$ (top) and filtered $\mathcal{A}'$ (bottom) along the branch surrounded by dashed contours in the tree. Right: extinction-based saliency map $\mathcal{M}_\mathcal{E}$ (top) and spotted meaningful objects surrounded by the colorized contour.

Fig. 6. Many of the local minima correspond actually to meaningless objects, and some local minima represent some regions that are very similar, only a representative one should be spotted.

We propose to apply a connected filter to the shape space to get rid of the spurious local minima of $\mathcal{A}$. More precisely, a pruning of the Min-tree of the shape space is applied, which is well-known as a local minima killer. Note that some local minima of the filtered attribute $\mathcal{A}'$ are flat zones of the shape space, which implies that some local minima of $\mathcal{A}'$ may contain several local minima of $\mathcal{A}$. We select the region having the smallest $\mathcal{A}$ as the representative one for the corresponding flat zone of local minima of $\mathcal{A}'$. An example of the scheme of this object segmentation method is depicted in Fig. 6.

4.3 Hierarchical segmentation based on extinction values

By expanding the idea of object spotting described in Section 4.2, if we increase the pruning force, more and more local minima of $\mathcal{A}$ will be filtered out or absorbed by the nearby local minima having a smaller attribute value. So less and less local minima are spotted as interesting objects. In this sense, each local minimum has a certain possibility to be spotted as an interesting object. This possibility can be measured by the notion of extinction values $\mathcal{E}$ [9] of local minima, which reveals their importance, so does the meaningfulness of the objects corresponding to the local minima.

Let $\mathcal{AA}^\uparrow$ be an increasing attribute on the Min-tree $\mathcal{TT}$ constructed from the shape space, when the pruning force of $\mathcal{TT}$ based on $\mathcal{AA}^\uparrow$ increases slightly, some minima merge with other minima being more significant. The extinction value $\mathcal{E}$ for a minimum m_i is defined as the maximal pruning force for which m_i does not yet merge with another minimum being more significant. More specifically, let $\prec$ be a strict total order on the set of minima $m_1 \prec m_2 \prec \cdots \prec m_n$ in a decreasing order of significance, such that $m_i \prec m_{i+1}$ whenever $\mathcal{A}(m_i) < \mathcal{A}(m_{i+1})$. Let CC be the lowest connected component that contains both m_i and a minimum m_j with $j < i$. The extinction value for the minimum m_i is defined as the difference between the attribute $\mathcal{AA}^\uparrow$ of CC and of m_i: $\mathcal{AA}^\uparrow(CC) - \mathcal{AA}^\uparrow(m_i)$. The Figure 7 shows an example of the extinction value for three minima based on the attribute $\mathcal{AA}^\uparrow$ being the current value of $\mathcal{A}$. The order is $A \prec C \prec B$, and B merges with C, C merges with A.

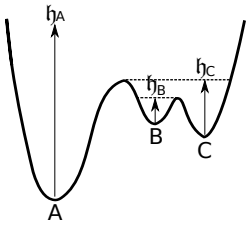


Fig. 7: Illustration of extinction values.

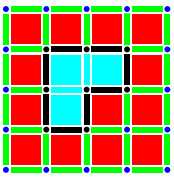


Fig. 8: Materialization of pixels with 0-faces (blue disks), 1-faces (green strips), and 2-faces (red squares). The original pixels are the 2-faces, the boundaries are materialized with 0-faces and 1-faces. The contour of the cyan region is composed of black 1-faces and 0-faces.

Instead of computing a single object spotting result as shown in Section 4.2, we compute a saliency map $\mathcal{M}_\mathcal{E}$ based on the extinction values of local minima. This map represents a hierarchy of object spotting results. More precisely, we weigh the extinction values $\mathcal{E}$ to the region boundaries of the corresponding local minima. To facilitate the representation of the boundaries, note that we use the Khalimsky's grid [38], where a region boundary is composed of a set of elements lying between pixels (2-faces) for a 2D image. The set of elements of a region boundary are materialized by 1-faces and 0-faces. A synthetic example is depicted in Figure 8.

The scheme of computation of the extinction-based saliency map $\mathcal{M}_\mathcal{E}$ relying on the use of Khalimsky's grid [38] is detailed below (an efficient computation algorithm is given in [5]):

1) Initialize the saliency map $\mathcal{M}_\mathcal{E}$ with 0; note that the size of $\mathcal{M}_\mathcal{E}$ is doubled compared to the original image f.

2) For each 1-face $e \in E$ in the saliency map, the value at this 1-face $\mathcal{M}_{\mathcal{E}}(e)$ is set to the maximal extinction value of the nodes representing the set of nested regions in the shape space having e as an element of its boundary. Note that, if there is no minimum among the set of nodes, we set $\mathcal{M}_{\mathcal{E}}(e)$ to 0.
3) For each 0-face o in the image, $\mathcal{M}_{\mathcal{E}}(o) = \max\{\, \mathcal{M}_{\mathcal{E}}(e) \,|\, e \text{ is a 1-face neighbor of } o \,\}$.

Each threshold of this saliency map $\mathcal{M}_{\mathcal{E}}$ represents an object spotting result, which corresponds to a result obtained by the method of object spotting described in Section 4.2 with a certain pruning force. This saliency map $\mathcal{M}_{\mathcal{E}}$ represents a hierarchy of segmentations based on the use of the shape space and an attribute $\mathcal{A}$. An example is given in Fig. 6. A threshold of the saliency map in Fig. 6 gives the four most meaningful objects whose boundaries are colorized with different colors.

5 Experimental results

As described in Section 4, there are two main ingredients in the proposed framework:

1) A shape space built from a hierarchical representation of the image containing the regions of interest.
2) An attribute characterizing each region, and its local minima represent the significant regions.

In this section, we show several experimental results of the proposed framework using different hierarchical representations and attributes. First of all, in Section 5.1, we use the α-tree to construct the shape space and an attribute $\mathcal{A}_f$ inspired from the work of [12] for generic image segmentation. Qualitative and quantitative results on BSDS300 [39] and BSDS500 [1] dataset shows that we obtain similar results with the original work in [12], but with the advantage of being hierarchical. At the same time, we also improve the hierarchy of constrained connectivity [14] by extending it with non-increasing attributes. Then in Section 5.2, we use the tree of shapes [22] to build the shape space and the attributes $\mathcal{A}_s$ relying only on shape information. Application of the shape-oriented saliency maps to traffic sign detection achieves results on par with the dedicated baseline algorithms on the German Traffic Sign Detection Benchmark (GTSDB) dataset [40]. Finally, we have experimented with a dedicated attribute characterizing the documents in Section 5.3 for document extraction in videos captured by smartphones. The shape space is again built from the tree of shapes. The resulting document-oriented saliency map achieves the first place in the Smartphone Document Capture and OCR (SmartDoc) competition organized at ICDAR 2015 [41].

5.1 Extending constrained connectivity

In this section, we use a binary Min-tree of the MST, which is equivalent to the α-tree (as proved in [15]) to construct the shape space. The images are regularized by anisotropic diffusion [42] in order to smooth their textures. The dissimilarity used for the MST is the maximal distance of the red, blue, and green channels taken independently.

We use an attribute $\mathcal{A}_f$ inspired from the work of Felzenszwalb and Huttenlocher [12]. In [12], the authors propose a region merging process that follows the edges of the MST by increasing order of the weights (dissimilarity). At the beginning, each flat zone is considered as an individual region. Then, when an edge (x, y) is considered, they search for the regions X and Y that respectively contain the points x and y. The regions X and Y are merged if

$$\text{Diff}(X,Y) \;<\; \min\{\, \text{Int}(X) + \frac{k}{|X|},\, \text{Int}(Y) + \frac{k}{|Y|} \,\}, \quad (1)$$

where $|\cdot|$ denotes the cardinality, $\text{Diff}(X,Y)$ is the minimum weight of the edge connecting the two components X and Y, $\text{Int}(X)$ is the largest weight in the MST of the region X, and k is a parameter favoring the merging of small regions (a larger k causes a preference for larger components). However, k is not a scale parameter in the sense of the *causality principle*: as shown in [43], a contour present at a scale k_1 is not always present at a scale $k_2 < k_1$. This is a practical difficulty for a user that wants to tune the approach to a particular task. By producing a hierarchy based on the merging criterion, we remove this difficulty.

The merging criterion defined by Eq. (1) depends on the parameter k at which the regions X and Y are observed. So let us consider the attribute $\mathcal{A}_f$ as the k given by $k = \max\{\, (\text{Diff}(X,Y) - \text{Int}(X)) \times |X|,\, (\text{Diff}(X,Y) - \text{Int}(Y)) \times |Y| \,\}$. That is to say, for each region R, let R_{c1} and R_{c2} be the two children of R in the binary Min-tree of MST, then the attribute $\mathcal{A}_f$ for region R is given by

$$\begin{aligned}\mathcal{A}_f(R) \;=\; \max\{\, &(\text{Diff}(R_{c1},R_{c2}) - \text{Int}(R_{c1})) \times |R_{c1}|, \\ &(\text{Diff}(R_{c1},R_{c2}) - \text{Int}(R_{c2})) \times |R_{c2}| \,\}. \quad (2)\end{aligned}$$

An example of the extinction-based saliency map $\mathcal{M}^0_{\mathcal{E}}$ using inverted $\mathcal{A}_f$ is illustrated in Fig. 9 (b) for the input image in Fig. 9 (a). Observe that in this saliency map, there are many small regions that are very salient, a step of refinement is required. As performed classically in mathematical morphology, we propose to apply a grain filter [44] followed by an ultrametric watershed [4] on the initial saliency map $\mathcal{M}^0_{\mathcal{E}}$. The interested reader can refer to [8, 4] and the Chapter 7 of [45] for more details. The final refined saliency map $\mathcal{M}_{\mathcal{E}}$ is illustrated in (c). One level of segmentation obtained by thresholding this final saliency map $\mathcal{M}_{\mathcal{E}}$ is shown in Fig. 9 (d).

Fig. 10 and Fig. 11 show the saliency maps computed on some images from the BSDS500 dataset [1], together with some segmentations extracted from the hierarchies. Two evaluation schemes are provided in [1]. In the first one, the same fixed threshold level (observation scale) is used for all saliency maps in the dataset; we refer to it as the optimal dataset scale (ODS). In the second one, we evaluate the performance using an image-dependent

(a) Input image (b) Initial saliency map $\mathcal{M}_{\mathcal{E}}^{0}$

(c) Final saliency map $\mathcal{M}_{\mathcal{E}}$ (d) One segmentation

Fig. 9: An example showing the scheme of the saliency map computation. The saliency maps are inverted for better visualization.

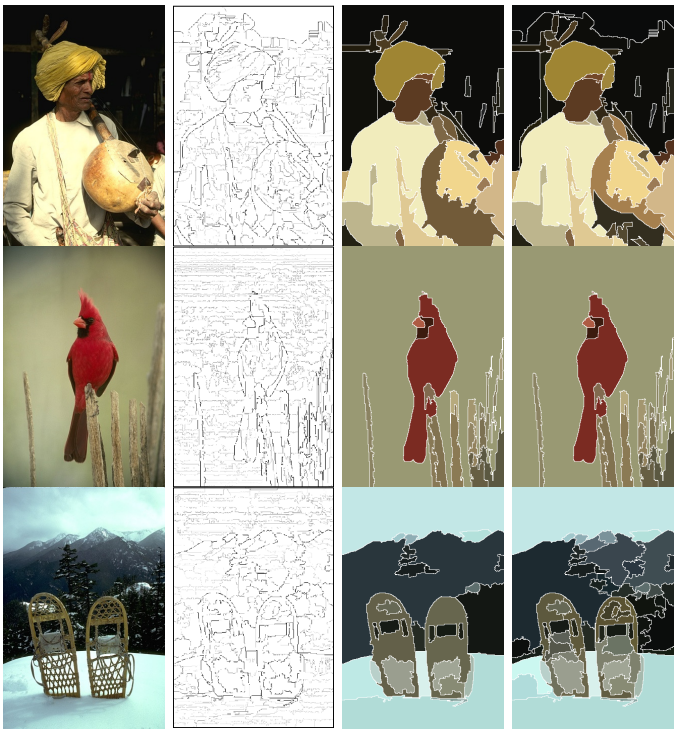


Fig. 10: Hierarchical segmentation results on the BSDS500. From left to right: Image, inverted saliency map, and segmentations at the optimal dataset scale (ODS) and optimal image scale (OIS).

threshold for each saliency map; we refer to this choice as the optimal image scale (OIS).

The quantitative evaluation is performed using the boundary-based precision-recall curves and the region-based performance measurements described in [1], in terms of Ground-Truth (GT) Covering criterion and Probabilistic Rand Index (PRI). A high GT covering and a large PRI are required for a "good" partition. The

Fig. 11: Additional hierarchical segmentation results on the BSDS500. From top to bottom: Image, inverted saliency map, and segmentations at the optimal dataset scale (ODS) and optimal image scale (OIS).

interested reader can refer to [1] for more details about these measures.

Here, we compare our results with three closely related work: 1) the original graph-based image segmentation in [12]; 2) another method of hierarchical graph-based image segmentation proposed by Guimarães *et al.* in [43], also relying on the same criterion popularized by [12]; and 3) the classical use of the α-tree by thresholding the hierarchy *w.r.t.* the local range α. The boundary-based precision-recall curves are illustrated in Fig. 12. The region-based comparison is given in Table 1. Our proposed extinction-based saliency map $\mathcal{M}_{\mathcal{E}}$ achieves a similar result as the original method in [12], but with the advantage that it produces a hierarchy. Note that the goal of this experiment is not to show that the proposed framework outperforms the state-of-the-art methods for generic segmentation. We choose to use a simple attribute in this experiment, which is a proof of concept for the proposed framework. We would expect learning the attributes to improve the performance. This is one of our future work.

Our method can also be seen as an extension of the hierarchy of constrained connectivity in the sense that we use a non-increasing attribute instead of an increasing one. As shown in Table 1, the proposed method improves the classical use of hierarchy of constrained connectivity.

5.2 Shape-oriented hierarchical segmentations

In the previous experiment, the attribute measures somehow the meaningfulness of each region in the shape space in the sense of generic image segmentation. We now demonstrate the interest of using an attribute char-

Method	BSDS300 [39]					BSDS500 [1]				
	GT Covering			Prob. Rand. Index		GT Covering			Prob. Rand. Index	
	ODS	OIS	Best	ODS	OIS	ODS	OIS	Best	ODS	OIS
gpb-owt-ucm [1]	0.59	0.65	0.75	0.81	0.85	0.59	0.65	0.74	0.83	0.86
Mean Shift [46]	0.54	0.58	0.66	0.78	0.80	0.54	0.58	0.66	0.79	0.81
FH [12]	0.51	0.58	0.68	0.77	0.82	0.52	0.57	0.69	0.80	0.82
Our	0.51	0.58	0.67	0.78	0.82	0.51	0.59	0.67	0.80	0.83
Guimarães [43]	-	-	-	-	-	0.46	0.53	0.60	0.76	0.81
α-tree [14, 11]	0.45	0.54	0.63	0.76	0.81	0.44	0.53	0.63	0.78	0.82
Ncuts [47]	0.44	0.53	0.66	0.75	0.79	0.45	0.53	0.67	0.78	0.80

TABLE 1: Region benchmarks on the BSDS300 [39] and BSDS500 [1]. Note that this experiment is just a proof of concept for the proposed framework. The goal is not to show that it outperforms the state-of-the-art methods for generic segmentation (see corresponding text for discussion).

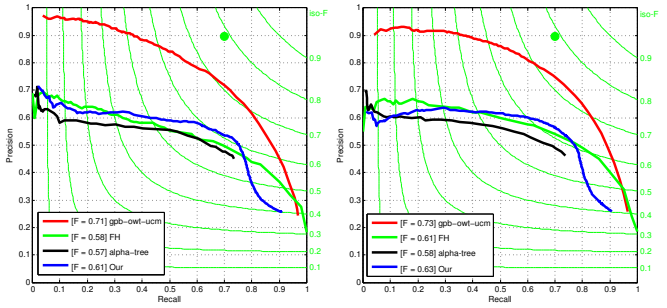


Fig. 12: Boundary benchmark on the BSDS300 [39] (left) and BSDS500 dataset [1] (right). Note that the goal of this experiment is not to show that the proposed framework outperforms the state-of-the-art methods for generic segmentation. This experiment is just a proof of concept for the proposed framework.

Method	Detection rate		Area under curve	
	Prohibitive	Danger	Prohibitive	Danger
Our	96%	**95%**	**92.16%**	**93.10%**
Viola-Jones [48]	**98.8**%	74.6 %	90.81%	46.26%
HOG+LDA [40]	91.3%	90.7%	70.33%	35.94%
Hough-like [49]	55.3%	65.1%	26.09%	30.41%

TABLE 2: Quantitative evaluation on GTSDB test dataset [40]. Note that the detection rate of the other methods is at a precision of 10% for both prohibitive and danger sign. Our depicted detection rate is at a precision of 59% (*resp.* 41%) for prohibitive (*resp.* danger) sign.

acterizing the shape of each region, which results in a shape-oriented hierarchical segmentations.

As examples, we use two shape-based attributes that describe the roundness denoted as $\mathcal{A}_\circ$ and respectively the upper triangularity denoted as $\mathcal{A}_\triangle$ of each shape. More precisely, the circularity $\mathcal{A}_\circ$ is given by the standard deviation of distances between the boundaries (composed of a set of 0-faces and 1-faces as illustrated in Fig. 8) of each region and the corresponding center. A round region would have a small value for $\mathcal{A}_\circ$. The upper triangularity $\mathcal{A}_\triangle$ is defined relying on the overlap between the region R and its best fit upper triangle $R_\triangle$ based on the Jaccard similarity coefficient:

$$\mathcal{A}_\triangle(R) = 1 - |R \cap R_\triangle|/|R \cup R_\triangle|, \qquad (3)$$

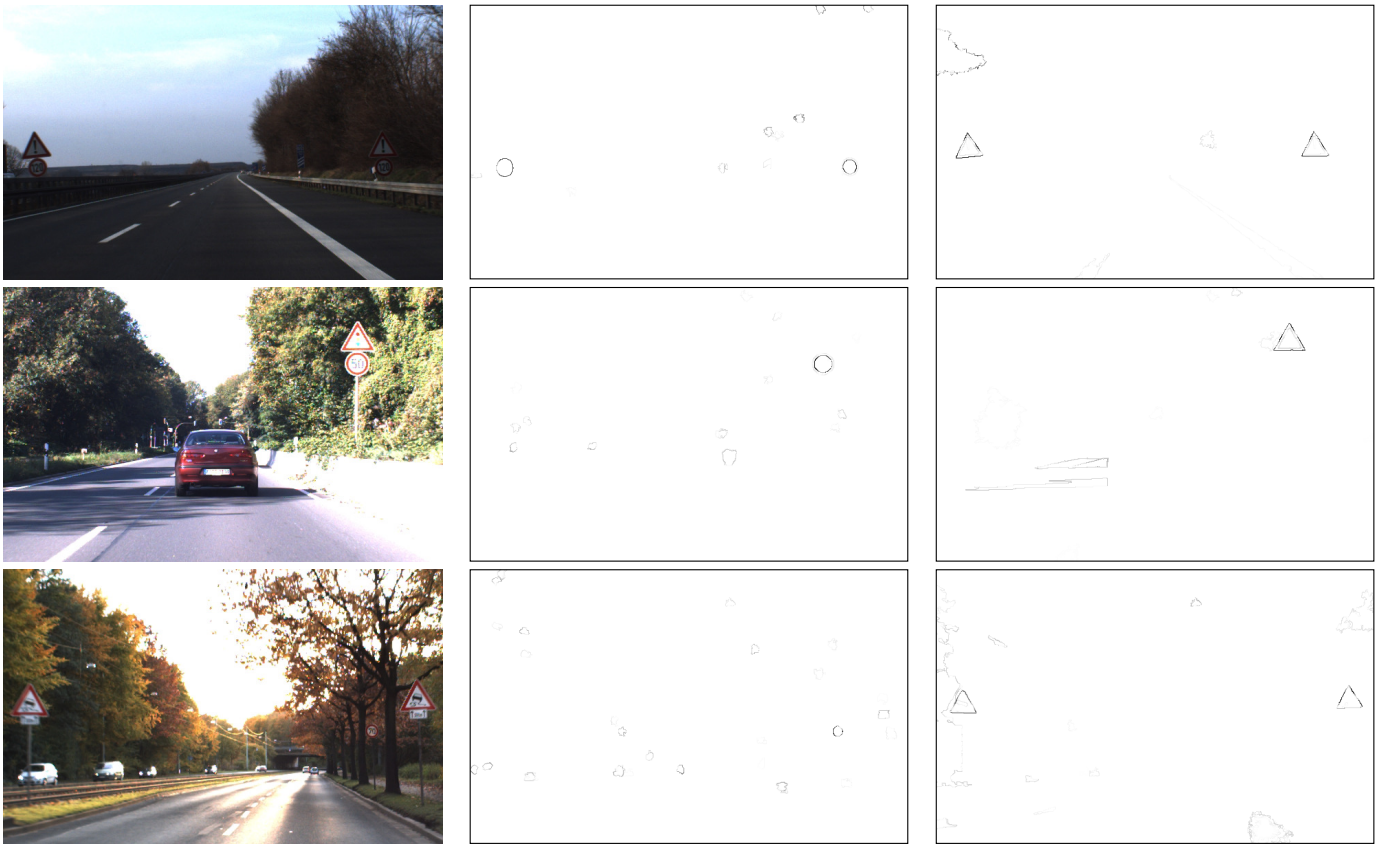
where $|\cdot|$ stands for the cardinality. For a given region R, its best fit upper triangle is determined by the three furthest point $p_u, p_{bl}, p_{br} \in R$ *w.r.t.* its center p_o, that are respectively above p_o, on the bottom left of p_o, and on the bottom right of p_o. An upper triangular region would have a small value for $\mathcal{A}_\triangle$. Note that these two attributes can be efficiently computed during the shape space construction [5].

We apply the proposed framework of hierarchical segmentations using the shape-based attributes $\mathcal{A}_\circ$ and $\mathcal{A}_\triangle$ to images of traffic scene, where the circular and upper triangular traffic signs are very common cases. We use the tree of shapes [22] to build the shape space. Note that the original color images are converted to grayscale images to compute the tree of shapes.

We have tested the method on the German Traffic Sign Detection Benchmark (GTSDB) dataset [40], where an on-line quantitative evaluation is available (http://benchmark.ini.rub.de). The dataset contains 900 images 1360×800 with dramatic illumination condition and size variations (see the input images illustrated in Fig. 13). The 900 images are split into 600 training images with ground truth, and 300 test images without making available the ground truth. Several qualitative illustrations of the proposed shape-oriented saliency maps on the test dataset are depicted in Fig. 13. Note that due to the large size of the input image and for visualization purpose, the images in this figure are the sub-parts of the original images that contain traffic signs. The saliency maps are inverted for better visualization.

For quantitative evaluation, we have benchmarked our shape-oriented saliency map for the detection of prohibitive (circular red-and-white) and danger (upper triangular red-and-white) signs using the accompanied on-line evaluation system (see [40] for details). For our shape-oriented saliency maps, we threshold them with a fixed value for the whole test dataset to achieve object segmentation. We only keep the detected circular objects that are round enough (*i.e.*, the relative $\mathcal{A}_\circ$ *w.r.t.* the major length of the best fit ellipse is small enough), and the triangular objects whose orientations are near horizontal.

The quantitative results as compared to some baseline methods that form the basis of many more evolved methods are depicted in Table 2. Note that the Viola-



Input image | Circular-object-oriented $\mathcal{M}_{\mathcal{E}}$ | Upper-triangular-object-oriented $\mathcal{M}_{\mathcal{E}}$

Fig. 13: Qualitative illustrations of shape-oriented saliency maps on some images from GTSDB test dataset [40].

Jones detector [48] based on Haar-like features [48] and the method "HOG+LDA" [40] based on HOG features [50] and linear discriminant analysis (LDA) [51] are supervised methods. The Hough-like method [49] makes use of shape information via Hough-like voting scheme and color information based on many heuristics. Our shape-oriented saliency map that uses only shape information achieves comparable detection rate *w.r.t.* to the dedicated methods, but at a significant higher precision (59% and respectively 41% instead of 10%). By integrating very simply the color information (*e.g.*, there must be enough red pixels around the boundaries, and there must be some white pixels inside the object) to validate the detected objects, we improve the precision from 59% to 96% (*resp.* 41% to 98%) while preserving their detection rate. This yields a better area under curve *w.r.t.* to the baseline methods. Note that among the submitted methods for this task, some dedicated learning-based methods achieve almost nearly perfect results. Although we have not experimented with a dedicated attribute for this application (contrary to the document extraction described in the following section), we would expect that a dedicated attribute (*e.g.*, a learned attribute based on color and shape features) can also achieve nearly perfect results for this specific application. In the following experiment in Section 5.3, we provide an example using a specifically dedicated attribute that outperforms the state-of-the-art methods for a target application.

5.3 Smartphone document capture

In this section, we experiment a dedicated attribute for a specific application: document detection in videos captured by smartphones. These documents have two main features: 1) They have quadrangular shape; 2) They contain text and/or graphics.

We use an attribute $\mathcal{A}_d$ to capture these two characteristics. More specifically, the quadrangularity $\mathcal{A}_{\square}$ is similar to the triangularity defined in Eq. (3), but it measures how closely the region R fits a quadrangle. Let $R_{\square}$ be the best fit quadrangle of the region R. The quadrangularity $\mathcal{A}_{\square}$ is given by:

$$\mathcal{A}_{\square}(R) = 1 - |R \cap R_{\square}|/|R \cup R_{\square}|, \tag{4}$$

where $|\cdot|$ denotes the cardinality. A quadrangular region would have a small value for $\mathcal{A}_{\square}$. We also expect the document to contain text and/or graphics, *i.e.*, contains many shapes inside R. This can be measured by:

$$\mathcal{A}_c(R) = 1 - \sum_{C \in \mathcal{L}_R} (d(C) - d(R))/|R|, \tag{5}$$

where $d(R)$ stands for the depth (starting from 0 for the root) of the region R, and $\mathcal{L}_R$ is the set of leaf nodes contained in the region R: $\mathcal{L}_R = \{C \in \mathbb{C} \,|\, C \subset$

R and C is a leaf}. Note that for accuracy of $\mathcal{A}_c$, we start by preprocessing the image with a grain filter to get rid of the natural image noise due to the sensor. Thus we only consider large enough regions as textual or graphical contents. A region that contains text and/or graphics, as opposed to an empty page, will have a small value for $\mathcal{A}_c$. Finally, the attribute $\mathcal{A}_d$ characterizing the document region R is defined as a simple combination of $\mathcal{A}_\square$ and $\mathcal{A}_c$ given by:

$$\mathcal{A}_d(R) = \mathcal{A}_\square(R) + \mathcal{A}_c(R). \quad (6)$$

A document region is expected to have a small value for $\mathcal{A}_d$.

We use the tree of shapes constructed on either the L^* or the b^* channel of each frame (converted in the $L^*a^*b^*$ color space). The choice of channel is automatically performed by comparing the smallest attribute $\mathcal{A}_d$ on either the L^* or b^* channel. For the first frame in each video, we select the most salient shape (*i.e.*, the shape having the highest extinction value). For the other frames, we select the 10 most salient shapes in the saliency map as candidate document regions, and order them w.r.t. the attribute $\mathcal{A}_d$. As we expect the camera to remain relative still, the distance between the corners of the document detected in the previous frame with the current candidate shapes is used to eliminate mis-detections. For the frame t, we get a family $\{R_t^1, \ldots, R_t^{10}\}$ of candidate shapes. Let R_{t-1}^* be the extracted document region in the previous frame $t-1$, then the shape R_t^* with the smallest attribute $\mathcal{A}_d$ that has a distance to R_{t-1}^* below a given threshold $d_{\max}$ is detected as the document region for the frame t. If no such shape R_t^* exists, the detection fails. The four corners of the extracted document region R_t^* are used to define the final detected document.

We have applied this method on the dataset of competition on Smartphone Document Capture and OCR (SmartDoc) organized at ICDAR 2015 [41]. The dataset has a total of 150 videos and 24889 frames. It is composed of six different types of documents that are recorded on five different backgrounds. The dataset covers a range of different types of documents in terms of textual and graphical contents, and a number of imaging conditions such as change of illuminations, change of perspectives, motion blurs, and partial occlusions and superposition of documents (see Fig. 14 for some illustrative frames).

Several qualitative results are depicted in Fig. 14. Due to the large size of the input frames, only sub-parts of the original frames are shown in this figure. For better visualization, the saliency maps are inverted. Note that a common misdetection of the proposed method lies on a sub-document retrieval, *i.e.*, with documents having graphical contents that may appear themselves as documents (*e.g.*, large tables). Indeed, a rectangular graphic or a table inside the document may have a lower attribute value $\mathcal{A}_d$ than the one of the expected whole document. For this challenge, we avoid this problem by using some prior knowledge of the document size (the candidate region could not be too small or too large). Note also that some videos have problems due to partial occlusions and superposition of documents. Such problems can be automatically detected because the attribute of the best detected shape is not small enough. In such cases, a preprocessing by opening and closing with a vertical structuring element is automatically applied to have a better shape space that contains the document region.

Ranking	Method	Jaccard Index	Confidence Interval
1	Our2	**0.9816**	**[0.9813, 0.9819]**
1	Our	**0.9716**	**[0.9710, 0.9721]**
2	ISPL-CVML	0.9658	[0.9649, 0.9667]
3	SmartEngines	0.9548	[0.9533, 0.9562]
4	NetEase	0.8820	[0.8790, 0.8850]
5	A2iA run 2	0.8090	[0.8049, 0.8132]
6	A2iA run 1	0.7788	[0.7745, 0.7831]
7	RPPDI-UPE	0.7408	[0.7359, 0.7456]
7	SEECS-NUST	0.7393	[0.7353, 0.7432]

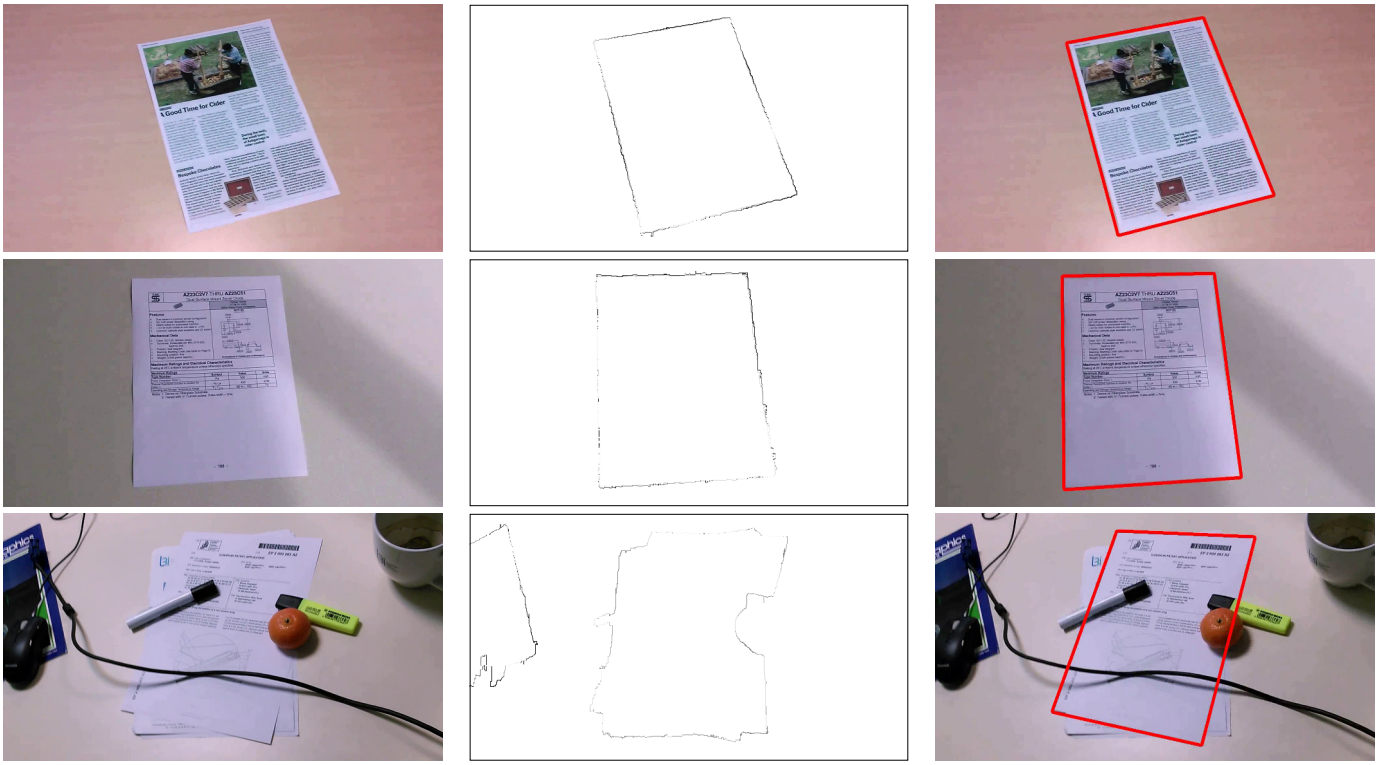
TABLE 3: Quantitative results on the dataset of SmartDoc competition [41]. Our result is further improved by adding a postprocessing for the videos with partial occlusions and superposition of documents.

Quantitative assessment is depicted in Table 3. The numerical evaluation is based on the Jaccard similarity coefficient between the ground truth defined by the four corners of the documents and the detected document regions. The 95% confidence intervals are also presented to demonstrate the robustness of different methods. As shown in Table 3, our method named "Our" was awarded the first place out of 7 participants in this challenge. We further improved our result on the challenge by applying a postprocessing for the videos automatically detected as suffering partial occlusions and superposition of documents. More precisely, we re-estimate the four corners of the detected shape such that the quadrangle defined by the four corners fits the best the contents inside the detected shape. This improved result, named "Our2", is also depicted in Table 3. Note that some videos are available as supplementary material accompanying this article.

6 Conclusion

In this paper, we have presented a general framework for transforming any hierarchical representation of an image into a hierarchy of segmentations highlighting specific objects. The two main components of the framework are the shape space construction and the attribute computation. More precisely, the framework relies on the shape space [6], a graph representation of a set of regions extracted from the original image. We weigh the shape space with an attribute $\mathcal{A}$ capturing the expected characteristics of regions of interest. The basic idea of the proposed framework is to consider all the local minima of the weighted shape space as candidate regions for a segmentation. Then the boundaries of these regions are weighed by the extinction values [9] of $\mathcal{A}$, which yields an extinction-based saliency map $\mathcal{M}_\mathcal{E}$.

A limitation of some previous work from the field of mathematical morphology is that the criterion used to



Input frame | Document-oriented $\mathcal{M}_{\mathcal{E}}$ | Extracted document

Fig. 14: Qualitative illustrations of the proposed method on some images from the dataset of SmartDoc competition [41]. Note that the full videos are available as supplementary material accompanying this article.

modify a hierarchy had to be *increasing* [37]. Our framework does not have this limitation since we can use any criterion, as it is depicted in the variety and the diversity of the presented examples. As compared to the classical state-of-the-art approaches that consist of cutting a hierarchy of segmentations [3, 2, 14, 4, 17, 18, 20, 19], each threshold of the extinction-based saliency map gives a segmentation result which is *a priori* different from any cut of the original hierarchy of segmentations. Actually, this process enlarges the set of possible partitions for a given hierarchy of segmentations. Besides, the proposed method allows us to obtain hierarchical segmentations from the threshold decomposition-based trees (*i.e.*, Min-tree, Max-tree, and tree of shapes), which are widely used in mathematical morphology and image processing. Furthermore, expected objects are brought to the fore in the saliency map.

The interest and the versatility of the proposed framework is demonstrated with three experiments, varying the input hierarchical representation to build the shape space and the attribute capturing the expected characteristics of different objects of interest. First, we have used the hierarchy of constrained connectivity to build the shape space. An attribute inspired by the work in [12] is used to characterize the meaningfulness of each region in the sense of generic image segmentation. Both qualitative and quantitative results on the BSDS500 dataset [1] show that the proposed framework improves on the hierarchy of constrained connectivity, and achieves results close to the ones of the original work in [12], but with the advantage of being hierarchical. Parameter tuning is easier as compared to the method in [12]. Secondly, we have applied the shape-oriented saliency map offered by the proposed framework to circular and upper triangular traffic sign detection. Quantitative results on the GTSDB test dataset [40] shows that using only shape information, the proposed framework achieves results comparable to the state-of-the-art baseline methods, that are supervised approaches. Last, we have used the proposed framework on preprocessed images with a specifically designed attribute to extract documents in videos, and participated to the ICDAR competition SmartDoc 2015. Our document-oriented saliency map method achieves the first place among the 7 participants. We further improved our result on this challenge by applying a dedicated postprocessing.

The proposed extinction-based saliency maps rely on a shape space and on an attribute. In future work, we would like to investigate some other trees to construct the shape space: the binary partition tree [29], the color tree of shapes [52], and the gbp-owt-ucm [1]. Indeed, it has been shown in [53] that the shape space built from gpb-owt-ucm is very useful for polyp segmentation, which is critical for colorectal cancer diagnosis. A good result is obtained in [53] by an ad-hoc method trying to select the most elliptical region from the hierarchy

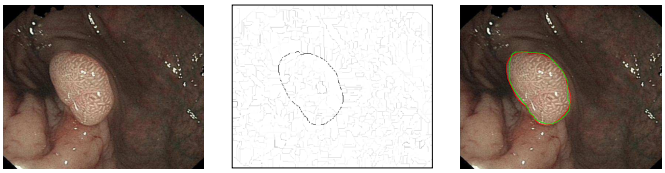


Fig. 15: Polyp segmentation using the proposed framework. Left: input image; Middle: elliptic-object-oriented saliency map $\mathcal{M}_\mathcal{E}$; Right: In green the manual annotation and in red the result by thresholding $\mathcal{M}_\mathcal{E}$.

of gpb-owt-ucm. This simple selection strategy can be extended using our elliptic-object-oriented saliency map. We would expect comparable performance, particularly for the case that more than one polyp are present in the image. An illustration is given in Fig. 15. We would also like to study some other shape spaces, in particular those which have a graph representation but which cannot be structured into a tree representation; it is for example the case of the component-graphs of multi-valued images [54], but also of any family of segmentations. Using attributes based on statistical measurements [55, 56] are also some interesting research perspectives. Last, but not the least, rather than using an *engineered* attribute, we envision that *learning* the attributes would bring many benefits to the practice.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments that greatly contributed to improve this paper. The authors are also grateful to Joseph Chazalon for providing the data and the evaluation system for smartphone document capture.

REFERENCES

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[2] L. Guigues, J. P. Cocquerez, and H. L. Men, "Scale-sets image analysis," *International Journal of Computer Vision*, vol. 68, no. 3, pp. 289–317, 2006.

[3] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 18, no. 12, pp. 1163–1173, 1996.

[4] L. Najman, "On the equivalence between hierarchical segmentations and ultrametric watersheds," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 3, pp. 231–247, 2011.

[5] Y. Xu, E. Carlinet, T. Géraud, and L. Najman, "Efficient computation of attributes and saliency maps on tree-based image representations," in *Proc. of Intl. Symp. on Mathematical Morphology*, ser. LNCS, vol. 9082. Springer, 2015, pp. 693–704.

[6] Y. Xu, T. Géraud, and L. Najman, "Connected filtering on tree-based shape-spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, to appear.

[7] P. Salembier, A. Oliveras, and L. Garrido, "Antiextensive connected operators for image and sequence processing," *IEEE Transactions on Image Processing*, vol. 7, no. 4, pp. 555–570, 1998.

[8] P. Salembier and M. H. F. Wilkinson, "Connected operators," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 136–157, 2009.

[9] C. Vachier and F. Meyer, "Extinction values: A new measurement of persistence," in *IEEE Workshop on Non Linear Signal/Image Processing*, 1995, pp. 254–257.

[10] Y. Xu, T. Géraud, and L. Najman, "Two applications of shape-based morphology: Blood vessels segmentation and a generalization of constrained connectivity," in *Proc. of International Symposium on Mathematical Morphology*, ser. LNCS, vol. 7883. Springer, 2013, pp. 390–401.

[11] G. Ouzounis and P. Soille, "Pattern spectra from partition pyramids and hierarchies," in *Proc. of Intl. Symp. on Mathematical Morphology*, ser. LNCS. Springer, 2011, vol. 6671, pp. 108–119.

[12] P. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, sep 2004.

[13] Y. Xu, T. Géraud, P. Monasse, and L. Najman, "Tree-based morse regions: A topological approach to local feature detection," *IEEE Trans. on Image Processing*, vol. 23, no. 12, pp. 5612–5625, 2014.

[14] P. Soille, "Constrained connectivity for hierarchical image partitioning and simplification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1132–1145, 2008.

[15] L. Najman, J. Cousty, and B. Perret, "Playing with Kruskal: Algorithms for morphological trees in edge-weighted graphs," in *Proc. of International Symposium on Mathematical Morphology*, ser. LNCS, vol. 7883. Springer, 2013, pp. 135–146.

[16] J. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," in *Proceedings of the American Mathematical Society*, vol. 7, no. 1, 1956, pp. 48–50.

[17] J. Serra, B. R. Kiran, and J. Cousty, "Hierarchies and climbing energies," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2012, pp. 821–828.

[18] J. Serra and B. R. Kiran, "Optima on hierarchies of partitions," in *Proc. of International Symposium on Mathematical Morphology*, ser. LNCS, vol. 7883. Springer, 2013, pp. 147–158.

[19] B. R. Kiran and J. Serra, "Global-local optimizations by hierarchical cuts and climbing energies," *Pattern Recognition*, vol. 47, no. 1, pp. 12–24, 2014.

[20] J. Cardelino, V. Caselles, M. Bertalmío, and G. Randall, "A contrario selection of optimal partitions for image segmentation," *SIAM J. on Imaging Sciences*, vol. 6, no. 3, pp. 1274–1317, 2013.

[21] A. Desolneux, L. Moisan, and J.-M. Morel, "Meaningful alignments," *Intl. J. of Computer Vision*, vol. 40, no. 1, pp. 7–23, 2000.

[22] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 860–872, 2000.

[23] V. Caselles, B. Coll, and J. Morel, "Topographic maps and local contrast changes in natural images," *International Journal of Computer Vision*, vol. 33, no. 1, pp. 5–27, 1999.

[24] J. J. Koenderink, "The structure of images," *Biological Cybernetics*, vol. 50, no. 5, pp. 363–370, 1984.

[25] F. Cao, P. Musé, and F. Sur, "Extracting meaningful curves from images," *J. of Math. Imaging and Vision*, vol. 22, pp. 159–181, 2005.

[26] E. R. Urbach, J. B. T. M. Roerdink, and M. H. F. Wilkinson, "Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 272–285, 2007.

[27] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *International Journal of Computer Vision*, vol. 88, no. 3, pp. 382–403, 2010.

[28] Y. Song and A. Zhang, "Analyzing scenery images by monotonic tree," *Multimedia Syst.*, vol. 8, no. 6, pp. 495–511, 2003.

[29] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation and information retrieval," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 561–576, 2000.

[30] M. H. F. Wilkinson and M. A. Westenberg, "Shape preserving filament enhancement filtering," in *Proc. of Intl. Conf. on Med. Image Computing and Comp.-Ass. Intervention*, 2001, pp. 770–777.

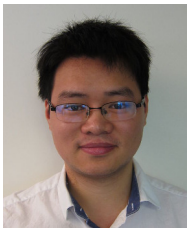
[31] C. Ballester, V. Caselles, L. Igual, and L. Garrido, "Level lines selection with variational models for segmentation and encoding," *Journal of Mathematical Imaging and Vision*, vol. 27, pp. 5–27, 2007.

[32] H. Lu, J. C. Woods, and M. Ghanbari, "Binary partition tree analysis based on region evolution and its application to tree simplification," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1131–1138, 2007.

[33] Y. Xu, T. Géraud, and L. Najman, "Salient level lines selection using the mumford-shah functional," in *Proc. of IEEE International Conference on Image Processing*, 2013, pp. 1227–1231.

[34] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE Transactions on Image Processing*, vol. 17, no. 11, pp. 2201–2216, 2008.

[35] Y. Xu, T. Géraud, and L. Najman, "Context-based energy estimator: Application to object segmentation on the tree of shapes," in *Proc. of IEEE Intl. Conf. on Image Processing*, 2012, pp. 1577–1580.
[36] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. of British Machine Vision Conference*, 2002, pp. 384–396.
[37] B. Perret, J. Cousty, J. C. R. Ura, and S. J. F. Guimarães, "Evaluation of morphological hierarchies for supervised segmentation," in *Proc. of International Symposium on Mathematical Morphology*, ser. LNCS, vol. 9082. Reykjavik, Iceland: Springer, 2015, pp. 39–50.
[38] E. Khalimsky, R. Kopperman, and P. R Meyer, "Computer graphics and connected topologies on finite ordered sets," *Topology and its Applications*, vol. 36, no. 1, pp. 1–17, 1990.
[39] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. of IEEE International Conference on Computer Vision*, vol. 2, July 2001, pp. 416–423.
[40] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in *International Joint Conference on Neural Networks*, no. 1288, 2013.
[41] J.-C. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. M. Luqman, M. Mehri, N. Nayef, J.-M. OGIER, S. Prum, and M. Rusinol, "ICDAR2015 competition on smartphone document capture and OCR (SmartDoc)," in *Proc. of International Conference on Document Analysis and Recognition*, 2015.
[42] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.
[43] S. J. F. Guimaraes, J. Cousty, Y. Kenmochi, and L. Najman, "An efficient hierarchical graph based image segmentation," in *14th International Workshop on Structural and Syntactic Pattern Recognition*, Hiroshima, Japan, 2012.
[44] V. Caselles and P. Monasse, "Grain filters," *Journal of Mathematical Imaging and Vision*, vol. 17, no. 3, pp. 249–270, 2002.
[45] L. Najman and H. Talbot, *Mathematical Morphology: From Theory to Applications*. ISTE-Wiley, Jun. 2010.
[46] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
[47] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 1124–1131.
[48] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2001.
[49] S. Houben, "A single target voting scheme for traffic sign detection," in *Proc. of IEEE Intelligent Vehicles Symposium*, 2011, pp. 124–129.
[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.
[51] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.
[52] E. Carlinet and T. Géraud, "MToS: A tree of shapes for multivariate images," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5330–5342, dec 2015.
[53] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2144–2151, 2012.
[54] N. Passat and B. Naegel, "Component-trees and multivalued images: Structural properties," *Journal of Mathematical Imaging and Vision*, vol. 49, no. 1, pp. 37–50, 2014.
[55] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *Comp. Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, 2001.
[56] F. Calderero and F. Marques, "Region merging techniques using information theory statistical measures," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1567–1586, 2010.



Yongchao Xu received the Engineering degree in electronics and embedded systems from Polytech Paris-Sud, Orsay, France, the master's degree in signal processing and image processing from the Université Paris-Sud, Orsay, in 2010, and the Ph.D. degree in image processing and computer vision from Université Paris-Est, France, in 2013. He is currently with the EPITA Research and Development Laboratory (LRDE), Paris, France, and LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, as a Post-Doctoral Fellow. His research interests include mathematical morphology, image segmentation, medical image analysis, and local feature detection.

Edwin Carlinet received the Ing. degree from EPITA, Paris, France, in 2011, a M.Sc. in applied mathematics for computer vision and machine learning from the École Normale Supérieure Cachan, in 2012, and the Ph.D. degree in image processing and computer vision from Université Paris-Est, France, in 2015. He is currently working at DxO, Paris, France. His research interests include bio-informatics mathematical morphology, and statistical learning.

Thierry Géraud received a Ph.D. degree in signal and image processing from Télécom ParisTech in 1997, and the Habilitation à Diriger les Recherches from Université Paris-Est in 2012. He is one of the main authors of the Olena platform, dedicated to image processing and available as free software under the GPL licence. His research interests include image processing, pattern recognition, software engineering, and object-oriented scientific computing. He is currently working at EPITA Research and Development Laboratory (LRDE), Paris, France.

Laurent Najman Laurent Najman received the Habilitation à Diriger les Recherches in 2006 from the University of Marne-la-Valle, a Ph.D. of applied mathematics from Paris-Dauphine University in 1994 with the highest honour (Félicitations du Jury) and an Ingénieur degree from the Ecole des Mines de Paris in 1991. After earning his engineering degree, he worked in the Central Research Laboratories of Thomson-CSF for three years, working on some problems of infrared image segmentation using mathematical morphology. He then joined a start-up company named Animation Science in 1995, as director of research and development. The technology of particle systems for computer graphics and scientific visualisation, developed by the company under his technical leadership received several awards, including the European Information Technology Prize 1997 awarded by the European Commission (Esprit programme) and by the European Council for Applied Science and Engineering and the Hottest Products of the Year 1996 awarded by the Computer Graphics World journal. In 1998, he joined OC Print Logic Technologies, as senior scientist. He worked there on various problem of image analysis dedicated to scanning and printing. After ten years of research work on image processing and computer graphics problems in several industrial companies, he joined the Informatics Department of ESIEE, Paris in 2002, where he is a professor and a member of the Laboratoire dInformatique Gaspard Monge, Université Paris-Est Marne-la-Vallée. His current research interest is discrete mathematical morphology and discrete optimization.